

On the Convergence Properties of Optimal AdaBoost *

Joshua Belanich

Luis E. Ortiz

Department of Computer Science

Stony Brook University

Stony Brook, NY 11794-4410, USA

JOSHUA.BELANICH@STONYBROOK.EDU

LEORTIZ@CS.STONYBROOK.EDU

Abstract

In this paper, we establish the convergence of the Optimal AdaBoost classifier under mild conditions. We frame AdaBoost as a dynamical system, and provide sufficient conditions for the existence of an invariant measure. Employing tools from ergodic theory, we show that the margin for every example converges. More generally, we prove that the time average of any function of the weights over the examples converges. If the weak learner satisfies some common conditions, the generalization error is not changing much in the limit. We conjecture that these conditions are satisfied on almost every dataset, and show preliminary empirical evidence in support of that conjecture.

Keywords: AdaBoost, convergence dynamics, generalization error, margins, boosting

1. Introduction

Leo Breiman once called AdaBoost (Freund and Schapire, 1997) the best off-the-shelf classifier for a wide variety of datasets (Breiman, 1999). Fourteen years later, AdaBoost is still widely used because of its simplicity, speed, and theoretical guarantees for good performance. However, despite its overwhelming popularity, there is still a mystery surrounding its generalization performance (Mease and Wyner, 2008).

On each iteration of AdaBoost a new hypothesis, generated by a weak learning algorithm, is added to a running linear combination of hypotheses. Intuitively this combination of hypotheses is increasing in complexity the longer the algorithm is run. Meanwhile, the generalization performance of this ensemble tends to improve or remain stationary after a large number of iterations, contradicting standard VC-Dimension based bounds (Freund and Schapire, 1996; Drucker and Cortes, 1995; Breiman, 1998; Quinlan, 1996). In some cases, the generalization error continues to decrease long after the training error of the chain has reached zero (Schapire et al., 1998). A common graph depicting this behavior is Figure 1. Remarkably, the complicated combination of 1000 trees generalizes better than the simpler combination of 10.

Solving this paradox has been a driving force behind boosting research, and various explanations have been proposed. By far the most popular among them is the theory of margins. The generalization error of any convex combination of functions can be bounded by a function of their *margins* on the training examples, independent of the number of classifiers

*. Original: December 4, 2012

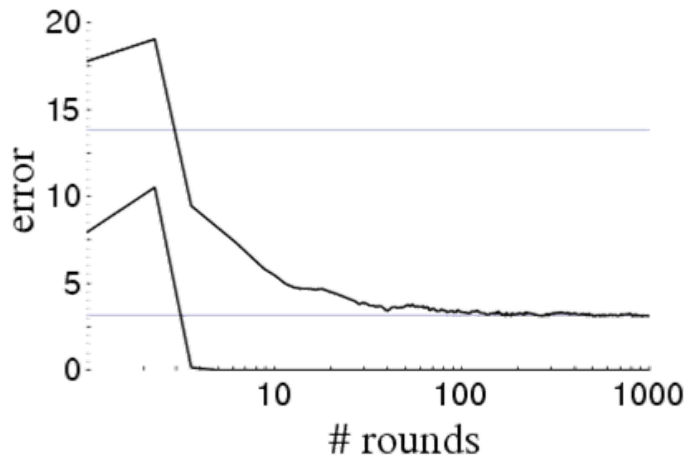


Figure 1: Training and test error (y-axis) for AdaBoosting C4.5 on the letter dataset for up to 1000 rounds (x-axis, in log-scale). This plot, which originally appeared in Schapire et al. (1998), is still featured in many tutorials and talks, but without any definitive formal explanation (see Breiman (1999) and Grove and Schuurmans (1998) for experimental evidence against the “max-margin theory” originally put forward as an explanation for this behavior by Schapire et al. (1998).) We believe our contribution in this paper provides a rather convincing, formal explanation for this behavior.

in the ensemble. AdaBoost provably produces large margins, and tends to continue to improve the margins after the training error has reached zero (Schapire et al., 1998).

The margin theory is effective at explaining AdaBoost’s generalization performance at a high level. But it still has its downsides. There is evidence both for and against the power of the margin theory to predict the quality of the generalization performance (Breiman, 1999; Rudin et al., 2004, 2007). But the most striking problem is that the margin bound is very loose: it does not explain the precise behavior of the error. For example, when looking at Figure 1 a couple of questions arise. Why is the generalization error not fluctuating wildly underneath the bound induced by the margin? Or even, why is the generalization error not approaching the bound? Remarkably, the error does neither of these things, and seems to converge.

This phenomenon is not unique to this dataset. We can see this convergence on many different datasets, both natural and synthetic. Even in cases where AdaBoost seems to be overfitting, the generalization performance tends to converge eventually. Take for example Figure 2. For the first 5000 rounds it appears that the algorithm is overfitting. Afterwards, its generalization error converges.

To account for this convergent behavior, Breiman (2001) conjectured that AdaBoost was an ergodic dynamical system. He argued that if this was the case, then the dynamics

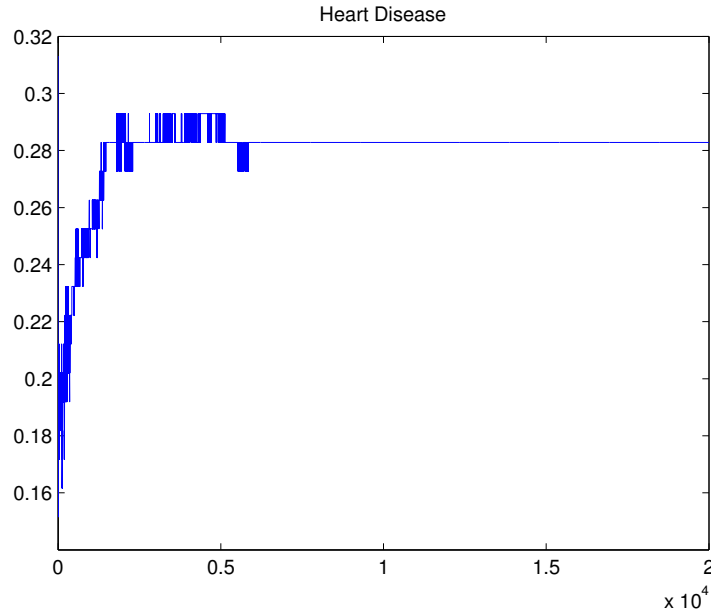


Figure 2: Test error (y-axis) of AdaBoosting decision-stumps on the Heart-Disease dataset (Frank and Asuncion, 2010b) for up to 20,000 rounds (x-axis). Note the same converging behavior exhibited for the letter data set in Figure 1. This converging behavior is typically observed in empirical studies of AdaBoost.

of the weights over the examples behaves like selecting from some probability distribution. Therefore AdaBoost can be treated as a random forest. Using the strong law of large numbers, it follows that the generalization error of AdaBoost converges for certain weak learners.

In this paper, we follow a similar approach. We frame AdaBoost as a dynamical system (Rudin et al., 2004). From here we establish sufficient conditions for an invariant measure on this dynamical system. We do not require this measure to be ergodic, so this is weaker than Breiman’s requirement. Using tools from ergodic theory, we show that such a measure implies the convergence of the time average of any function of the weights over the examples. In particular, this shows that the margin for every example converges. We also show that the AdaBoost classifier itself is converging if the weak learner satisfies certain conditions. Ultimately we prove our main result: the generalization error does not change much in the limit. We hope these results shed light on “the most important open problem in machine learning” (Breiman, 2002).

2. Related Work

Despite the flurry of convergence results for AdaBoost within the last 10 years, we believe our result is unique. We refer the reader to Schapire and Freund (2012) for a textbook account of the state-of-the-art in AdaBoost research.

In this section, we briefly discuss and contrast those contributions we believe are closest to ours of which we are aware. It is important to note that a number of these convergence results concern *variants* of AdaBoost, such as regularized boosting (see, e.g., Rosset et al. (2004), Lozano et al. (2006), Xi et al. (2009), and the references therein). Those convergence results that do concern AdaBoost in its original form show convergence for other aspects of the algorithm like the exponential loss, as discussed in the next paragraph. We will now walk through some of the recent research of AdaBoost.

A bulk of the asymptotic analysis on AdaBoost has been focused how it minimizes the exponential loss. AdaBoost can be viewed as a coordinate descent algorithm that iteratively minimizes the exponential loss (Breiman, 1999; Mason et al., 2000; Friedman et al., 1998). Under the weak learning condition, this minimization procedure was well understood, and has a quick rate. Later AdaBoost was shown to minimize the exponential loss even without the weak learning assumption (Collins et al., 2002; Zhang and Yu, 2005), but no rates were provided. Finally, Mukherjee et al. (2011) proved that AdaBoost enjoys a rate polynomial in $1/\epsilon$. Telgarsky (2012) achieves a similar result by exploring the primal dual relationship implicit in AdaBoost. These results all concerned the convergence of the *exponential loss*. Meanwhile, in this paper we are interested in the convergence of the basic Optimal AdaBoost classifier itself, along with its margins, and any time averaged function of its weights.

Consistency is another asymptotic concern about the behavior of AdaBoost. There are a number of papers that show that variants of AdaBoost are consistent (Zhang, 2004; Lugosi and Vayatis, 2004; Zhang and Yu, 2005), or that AdaBoost is consistent under certain assumptions (Peter J. Bickel, 2006). Bartlett and Traskin (2007) shows that AdaBoost is consistent if stopped at time $n^{1-\epsilon}$ for $\epsilon \in (0, 1)$, where n is the number of examples in the training set. Consistency is distinct from the notion of convergence in this paper. An algorithm is consistent if its generalization error approaches the Bayes risk in the limit of the number of examples in the training set. Here our concern is the generalization error in the limit of the number of iterations of the algorithm on a *fixed* sample.

Merler et al. (2007) provides empirical evidence for the statistical regularity underlying AdaBoost.

Others have also approached the study of AdaBoost from a dynamical-system perspective. Rudin et al. (2004) pioneered this approach, demonstrating that AdaBoost enters cycles in many low dimensional cases. They proved that when AdaBoost cycles, its asymptotic behavior can be fully understood: the minimum margin converges, and many of the results of this paper can be demonstrated. However, little is understood in the non-cyclic case. Chaotic non-cyclic behavior occurs on some higher dimensional cases: typical behavior for AdaBoost on large real world datasets. In fact, whether AdaBoost always cycles remains an open question to this date (Rudin et al., 2012).

We view our paper as an extension to the cyclic behavior of AdaBoost, as many of the results are the same, but our work encompasses non-cyclic, chaotic behavior.

3. Background and Notation

We make the assumption that all samples are taken from a probability space. The set of possible examples will be denoted $D = X \times \{0, 1\}$, where X are the instances and $\{0, 1\}$ are their labels. We will view D as part of a probability space, denoted $\mathbf{D} = (D, \Sigma, P)$. Σ is

the set of possible events, and P is a probability measure mapping $\Sigma \rightarrow \mathbf{R}$. Samples taken from \mathbf{D} will be denoted S , with $S \subset X$ and $S = \{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$. The set of weak hypothesis that AdaBoost will be provided is denoted by \mathcal{H} , and may be finite or infinite. This set of hypothesis induces a set of dichotomies on a training set S . We denote this set of dichotomies as $\text{Dich}(\mathcal{H}, S) = \{h^{(1)}, h^{(2)}, \dots, h^{(n)}\}$. An element h of $\text{Dich}(\mathcal{H}, S)$ is treated as a 0-1 row vector, with $h^{(i)}(j)$ being the label hypothesis $h^{(i)}$ gives example $x^{(j)}$. From this set we can derive an error matrix. This matrix has the form $M(i, j) = \mathbf{1}[y(j) = h^{(i)}(j)]$. A row i of the matrix can be thought of as a 0-1 vector indicating where dichotomy i is incorrect on the set S . We will abuse notation and treat M as a set. When we do this, the set elements are the rows of the matrix, we will call these rows dichotomies, and we denote them by η . For example, we often state “for all $\eta \in M$ ”. By this we mean “all rows η in M ”.

4. AdaBoost as a Dynamical System

This paper studies optimal AdaBoost as a dynamical system of the weights over the examples, in a way similar to Rudin et al. (2004). Here we break down the components of the AdaBoost algorithm and frame it as such a system. We will fix \mathcal{H} and S , therefore fixing $\text{Dich}(\mathcal{H}, S)$ and M . The state space of the system is the standard m -simplex, denoted

$$\Delta_m = \left\{ w \in \mathbf{R}^m \mid \sum_{i=1}^m w(i) = 1 \text{ and for all } i, w(i) \geq 0 \right\}.$$

We will often denote elements of Δ_m as w .

AdaBoost extensively uses the weighted error of a hypothesis in its weight update. The typical notion for this is

$$\text{err}(h, w) = \sum_{i=1}^m w(i) \mathbf{1}[h(x_i) \neq y_i].$$

However, for much of our analysis we will reduce AdaBoost to only using the rows of M in its weight update. When $\eta \in M$, we denote the error as

$$\text{err}(\eta, w) = \eta \cdot w.$$

As we are working with optimal AdaBoost, we need a notion of “best” row in M for any weight $w \in \Delta_m$. The standard notion for best row is any in $\arg \min_{\eta \in M} \eta \cdot w$. However, multiple rows may be in that set. We need a policy for how we break ties between rows with the lowest error, so we assume there exists such a tie-breaking function $\text{AdaSelect} : \mathcal{P}(M) \rightarrow M$.

Definition 1 *Given a weight $w \in \Delta_m$, our notion for best row in M for w is defined as*

$$\eta^w \equiv \text{AdaSelect} \left(\arg \min_{\eta \in M} \eta \cdot w \right).$$

This selection procedure naturally partitions Δ_m into regions where different rows are best, in the sense that they would be selected by AdaSelect.

Definition 2 *Given some arbitrary $\eta \in M$, define*

$$\sigma^*(\eta) \equiv \{w \in \Delta_m \mid \eta = \eta^w\}.$$

Note that $\sigma^*(\eta)$ may be open or closed for different η 's, depending on how ties are broken using AdaSelect. The closure of this set will also play an important role.

Definition 3 *Given some arbitrary $\eta \in M$, define*

$$\sigma(\eta) \equiv \left\{ w \in \Delta_m \mid \eta \in \arg \min_{\eta' \in M} \eta' \cdot w \right\}.$$

The set $\sigma(\eta)$, being the closure of $\sigma^*(\eta)$, is naturally closed. However, these sets no longer form a partition on Δ_m . Given two distinct dichotomies $\eta_1, \eta_2 \in M$, it is possible that $\sigma(\eta_1) \cap \sigma(\eta_2) \neq \emptyset$.

Sometimes it is convenient to consider only the subset of Δ_m where every row in M has non-zero error.

Definition 4 *The set of all weights in Δ_m with non-zero error on all dichotomies is defined as*

$$\sigma_0 \equiv \{w \in \Delta_m \mid \eta \cdot w > 0 \text{ for all } \eta \in M\}.$$

We will depart from standard notation for the AdaBoost weight update.¹ The notation we use will be more convenient for the main proofs in this paper. First, we have a notion of a *hypothetical* weight update. That is, given $w \in \Delta_m$ if we *assume* that $\eta = \eta^w$, where would the AdaBoost weight update take w ?

Definition 5 *Given an arbitrary row $\eta \in M$, we define $\mathcal{T}_\eta : \Delta_m \rightarrow \Delta_m$ component-wise as*

$$[\mathcal{T}_\eta(w)]_i \equiv \frac{1}{2} w(i) \times \left(\frac{1}{\eta \cdot w} \right)^{\eta_i} \left(\frac{1}{1 - \eta \cdot w} \right)^{1 - \eta_i}.$$

\mathcal{T}_η certainly does not trace out the actual trajectory of the AdaBoost weights. The true update first finds the best row η^w , and then applies $\mathcal{T}_{\eta^w}(w)$.

Definition 6 *The AdaBoost weight update is $\mathcal{A} : \Delta_m \rightarrow \Delta_m$, defined as*

$$\mathcal{A}(w) \equiv \mathcal{T}_{\eta^w}(w).$$

1. The form of the update we use has previously appeared in Grove and Schuurmans (1998); Oza (2001); Rudin et al. (2004).

We can now trace the trajectory of the AdaBoost weights by repeatedly applying \mathcal{A} to an initial weight picked within Δ_m . More specifically, if $w_1 \in \Delta_m$ is taken as some initial point, we can rederive any w_t in our original formulation of the algorithm with

$$w_t = \mathcal{A}^{(t-1)}(w_1) \quad (1)$$

where $\mathcal{A}^{(t-1)}$ denotes composing \mathcal{A} with itself $t - 1$ times.

We can also derive many of parameters calculated by AdaBoost solely in terms of w_t .

Definition 7 *The following are functions defined from Δ_m to \mathbf{R} .*

1. $\epsilon(w) \equiv \min_{\eta \in M} \eta \cdot w$
2. $\alpha(w) \equiv \frac{1}{2} \log \left(\frac{1 - \epsilon(w)}{\epsilon(w)} \right)$
3. $\chi_{\sigma^*(\eta)}(w) \equiv \mathbf{1}[w \in \sigma^*(\eta)]$

Fact 8 *The following equalities hold*

1. $\epsilon_t = \epsilon(w_t) = \epsilon(\mathcal{A}^{(t-1)}(w_1))$
2. $\alpha_t = \alpha(w_t) = \alpha(\mathcal{A}^{(t-1)}(w_1))$.
3. $\eta_t = \eta^{w_t} = \eta^{\mathcal{A}^{(t-1)}(w_1)}$

The parameters described in Fact 8 will be called secondary parameters, because they can be derived solely from the weight trajectory. We seek to understand the statistical convergence properties of these secondary parameters, and the properties of the mapping \mathcal{A} that causes such behavior.

5. Convergence of the AdaBoost Classifier

As mentioned at the end of the previous section, the secondary parameters of AdaBoost can be written as functions based solely on the trajectory of \mathcal{A} applied to some initial $w_1 \in \Delta_m$. Evidence suggests that not only are these averages converging, but the AdaBoost classifier itself is converging. The study of the convergence of the classifier, and its implications, is the main goal of this section.

Recall that AdaBoost classifies examples by computing

$$\text{sign}(F_T(x)) = \text{sign} \left(\sum_{t=1}^T \alpha_t h_t(x) \right).$$

Carefully note that in this case, h_t correspond to hypotheses in \mathcal{H} , not dichotomies in M . Certainly for $x \in S$, $F_T(x)$ does not converge as T approaches infinity (in fact, it must be growing at least linearly in T if the weak-learning assumption holds). So what do we mean by convergence of the AdaBoost classifier? We can replace $\text{sign}(F_T(x))$ with $\text{sign}(\frac{1}{T}F_T(x))$ or $\text{sign}(\text{margin}_T(x))$ in our notion of classification. Then, under certain assumptions, if

either $\frac{1}{T}F_T(x)$ or $\text{margin}_T(x)$ converge for all x , so does $\text{sign}(F_T(x))$ and $\text{sign}(\text{margin}_T(x))$ respectively. It is this way that we mean “convergence of the AdaBoost classifier”.

Using these results we can establish, under some reasonable conditions, convergence of the same functions on any x outside of the training set. The upshot is that given these kinds of convergence, we can say something strong about how the generalization error of the AdaBoost classifier behaves in the limit. Intuitively, if the AdaBoost classifier is effectively converging, so should its generalization error. This outlines the main contribution of this paper.

Crucial to our understanding of this convergence is Birkhoff’s Ergodic Theorem, stated as Theorem 9 below. This theorem gives us sufficient conditions for statistical convergence, which we will then apply to our secondary parameters. Taking center stage in this theorem is the notion of a *measure* and a *measure-preserving* dynamical system. To be able to apply Birkhoff’s Ergodic Theorem, we need to show the existence of some *measure* m such that $(\Delta_m, \mathbf{B}, \mathcal{A}, m)$ is a *measure-preserving* dynamical system. Please see the appendix for a closer look at these topics. The existence of such a measure is given in Proposition 14, but relies on Assumption 12. The context surrounding these is discussed in greater detail shortly.

Equally important in Birkhoff’s Ergodic Theorem is the notion of integrability, captured by the notation $f \in L^1(m)$. This notation says that f is integrable with respect to the measure m . The precise meaning of this is that, first and foremost, f is *measurable*. Second, that

$$\int |f| dm < \infty.$$

If these two conditions hold, it follows that $f \in L^1(m)$. Proposition 15 shows us that various parameters of AdaBoost are in $L^1(m)$, therefore can be analyzed using Theorem 9.

We are now ready to introduce the theorem.

Theorem 9 (*Birkhoff Ergodic Theorem*) Suppose $\mathcal{A} : X \rightarrow X$ is measure-preserving and $f \in L^1(m)$ for some measure m . Then $\frac{1}{n} \sum_{i=0}^{n-1} f(\mathcal{A}^i(x))$ converges almost everywhere to a function $f^* \in L^1(m)$. Also $f^* \circ \mathcal{A} = f^*$ almost everywhere and if $m(X) < \infty$, then $\int f^* dm = \int f dm$.

We care about the asymptotic behavior of AdaBoost, and want to disregard any of its transient states. Therefore we would like to look at a subset of its state space that the dynamics will limit towards, or stay within. The following set is, intuitively, the set of non-transient states of \mathcal{A} . More specifically, it is the set of states that AdaBoost can reach for any time step t .

Definition 10 $\Omega_\infty = \bigcap_{t=1}^\infty \mathcal{A}^{(t)}(\Delta_m)$

The set Ω_∞ can be thought of as a trapped attracting set in the typical sense, because Δ_m is compact and $\mathcal{A}(\Delta_m) \subseteq \text{Int}(\Delta_m)$. The continuity properties of AdaBoost on this set is important to establish an invariant measure in Section 7. But it is difficult to say anything important about this set yet.

It turns out that there are discontinuities at many points in the state space. It is not difficult to see that any point $w \in \Delta_m$ that yields more than one row in $\arg \min_{\eta \in M} \eta \cdot w$ will be a discontinuity. We will call this a type 1 discontinuity. Similarly, any point that has $\eta^w \cdot w = 0$ will also be a discontinuity, which we will call a type 2 discontinuity. In the following theorem, we establish that \mathcal{A} will be continuous on any point besides those just mentioned.

Theorem 11 *AdaBoost is continuous on all points w such that $w \in \text{Int}(\sigma(\eta)) \cap \sigma_0$ for some η .*

Proof Let $D \equiv \text{Int}(\sigma(\eta)) \cap \sigma_0$. Take any $w \in D$, and let $\{w_i\}$ be an arbitrary sequence such that $\lim_{i \rightarrow \infty} w_i = w$. Let $\{w_j\}$ be the tail of $\{w_i\}$ that is contained within D . Then, we have the following for all w_j

$$\mathcal{A}(w_j) = \mathcal{T}_\eta(w_j). \quad (2)$$

From Definition 5, for all w_j it follows that

$$[\mathcal{T}_\eta(w_j)]_n = w_j(n) \times \left(\frac{1}{\eta \cdot w} \right)^{\eta(n)} \left(\frac{1}{1 - (\eta \cdot w)} \right)^{1 - \eta(n)}.$$

Because $\lim_{j \rightarrow \infty} w_j = w$, we have $\lim_{j \rightarrow \infty} w_j(n) = w(n)$. Similarly, we have $\lim_{j \rightarrow \infty} \eta \cdot w_j = \eta \cdot w$. Furthermore, by the weak learning assumption we have $0 < \eta \cdot w < \frac{1}{2}$. Combining these facts, we see that

$$\lim_{j \rightarrow \infty} \mathcal{T}_\eta(w_j) = \mathcal{T}_\eta(w).$$

Recalling Equation 2, we complete the proof. ■

In the next section, we will prove that the weight trajectories of AdaBoost are eventually bounded away from type 2 discontinuities. Then we assume that AdaBoost satisfies the condition that it is bounded away from type 1 discontinuities. This assumption will be instrumental in our analysis, and gives us a way of proving the essential existence of an invariant measure. The assumption is formalized below. Roughly speaking, this assumption essentially says that, after a sufficiently long number of rounds either (1) the dichotomy corresponding to the optimal weak hypothesis for a round is unique with respect to the weights at that round, or (2) the dichotomies corresponding to the hypothesis that are tied for optimal are essentially the same with respect to the weights in Ω_∞ .

Assumption 12 (Optimal AdaBoost Eventually Has No Ties.) *There exists a compact set G such that $\Omega_\infty \subseteq G$ and, given any pair $\eta, \eta' \in M$, we have either*

1. $\sigma(\eta) \cap \sigma(\eta') \cap G = \emptyset$; or
2. for all $w \in G$, $\sum_{i: \eta(i) \neq \eta'(i)} w(i) = 0$.

We conjecture that this assumption holds almost always in practice.

Conjecture 13 (No Ties) *Assumption 12 holds for Optimal AdaBoost, modulo minimal conditions on the weak-hypothesis spaces, the process generating the training data, and, indirectly, the dichotomies they induce.*

Note that part (2) of Assumption 13 allows us to reduce the set of dichotomies to only those that will never become effectively the same from the standpoint of Optimal AdaBoost when dealing with Ω_∞ .

We now declare two propositions that cover the conditions required to apply Theorem 9. For now, we take these propositions for granted. They will be proved and discussed in later sections.

Proposition 14 covers the first condition of Theorem 9, that \mathcal{A} is a measure-preserving dynamical system for some measure μ . Afterwards, Proposition 15 covers the second condition for the secondary parameters of AdaBoost that will be essential in the impending analysis. By combining the two propositions we are able to use Birkhoff's Ergodic Theorem as a hammer to prove our main results.

Proposition 14 *Suppose Assumption 12 holds. Then there exists a Borel probability measure μ on Δ_m such that $(\Delta_m, \mathcal{B}, \mu, \mathcal{A})$ is a measure-preserving dynamical system.*

Proposition 15 *Let μ be a Borel probability measure on Δ_m . Then the following functions are in $L^1(\mu)$:*

1. $\epsilon(w) = \min_{\eta \in M} \eta \cdot w$
2. $\alpha(w) = \frac{1}{2} \log \left(\frac{1 - \epsilon(w)}{\epsilon(w)} \right)$
3. $\chi_{\sigma^*(\eta)}(w) = \mathbf{1}[w \in \sigma^*(\eta)]$

If we take the above propositions for granted, we can get a convergence result for $\frac{1}{T} F_T(x_i)$ for x_i in the training set. Note that in the below theorem we depart from the standard notations of $F_T(x_i) = \sum_{t=1}^T \alpha_t h(x_i)$. The new notation defines $F_T(x_i)$ in terms of the rows in the matrix M constructed from the dichotomies of H . These dichotomies are defined over $\{0, 1\}$, unlike the hypotheses over $\{-1, 1\}$, so we need to scale and translate them appropriately. The new notation for $F_T(x_i)$ results in the exact same values as the one defined over hypothesis.

Theorem 16 *Let $F_T(x_i) : S \rightarrow \mathcal{R}$, written $F_T(x_i) = \sum_{t=1}^T 2\alpha_t(\eta_t(i) - 1)$, be the AdaBoost classifier function at round T . For all $x_i \in S$, the limit $\lim_{T \rightarrow \infty} \frac{1}{T} F_T(x_i)$ exists.*

Proof Let

$$C_\eta(T) \equiv \sum_{t=1}^T \alpha(w_t) \chi_{\sigma^*(\eta)}(w_t).$$

From our assumptions it is easy to deduce that $\alpha \cdot \chi_{\sigma^*(\eta)}$ is measurable. Furthermore, we have

$$\int_{w \in \Delta_m} \alpha(w) \chi_{\sigma^*(\eta)}(w) d\mu \leq \int_{w \in \Delta_m} \alpha(w) d\mu < \infty.$$

Whereby it follows that $\alpha \cdot \chi_{\sigma^*(\eta)}(w) \in L^1(\mu)$. Applying Birkhoff's Ergodic Theorem, we see that $\lim_{T \rightarrow \infty} \frac{1}{T} C_\eta(T)$ exists for all η . For convenience, set $C_\eta^* = \lim_{T \rightarrow \infty} \frac{1}{T} C_\eta(T)$.

We are restricting ourselves to the set $S \subset X$, hence we can write

$$\frac{1}{T} F_T(x) = \sum_{\eta \in M} \frac{1}{T} 2C_\eta(T)(\eta(i) - 1).$$

Finally, by taking limits we see that

$$\begin{aligned} \lim_{T \rightarrow \infty} \frac{1}{T} F_T(x) &= \sum_{h \in M} \lim_{T \rightarrow \infty} \left[\frac{1}{T} C_\eta(T) \right] 2(\eta(i) - 1) \\ &= \sum_{\eta \in M} 2C_\eta^*(\eta(x) - 1). \end{aligned}$$

■

As a corollary to this theorem, we can show convergence for the margin on any example in the training set.

Corollary 17 *Consider the function $\text{margin}_T(x_i) : S \rightarrow \mathcal{R}$*

$$\text{margin}_T(x_i) = F_T(x_i) \left(\sum_{t=1}^T \alpha_t \right)^{-1}.$$

For all $x \in S$, the limit $\lim_{T \rightarrow \infty} \text{margin}_T(x_i)$ exists.

Proof By Birkhoff's ergodic theorem, $\Theta = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \alpha(w_t)$ exists. From the weak learning assumption, we know that $\epsilon_t = \epsilon(w_t) < \frac{1}{2} - \gamma$ for some $\gamma > 0$. This gives us a lower bound $\alpha_* > 0$ on $\alpha(w_t)$. Using this lower bound, we see that

$$\begin{aligned} \Theta &= \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \alpha(w_t) \\ &\geq \lim_{T \rightarrow \infty} \frac{1}{T} (\alpha_* T) \\ &\geq \alpha_* \\ &> 0. \end{aligned}$$

Now, we can say that $\lim_{T \rightarrow \infty} T \left(\sum_{t=1}^T \alpha_t \right)^{-1} = \frac{1}{\Theta}$. Combining this with Theorem 16, we have for all $x_i \in S$

$$\begin{aligned} \lim_{T \rightarrow \infty} \text{margin}_T(x_i) &= \lim_{T \rightarrow \infty} F_T(x_i) \left(\sum_{t=1}^T \alpha_t \right)^{-1} \\ &= \lim_{T \rightarrow \infty} \frac{F_T(x_i)}{T} \left(\sum_{t=1}^T \alpha_t \right)^{-1} T \\ &= \frac{1}{\Theta} \sum_{\eta \in M} 2C_\eta^*(\eta(i) - 1) \\ &= \sum_{\eta \in M} 2\tilde{C}_\eta(\eta(i) - 1). \end{aligned}$$

Where \tilde{C}_η is a probability distribution over the rows in M . ■

We run into some difficulties if we try to extend the above results to the whole instance space X instead of just the training set S . On S , we know that $2(\eta(i) - 1)$ will correspond directly with some $h(x_i)$. However, outside of S our η 's are no longer defined, because they are simply 0-1 vectors over the examples in S . To evaluate $F_T(x)$ for an arbitrary $x \in X$, we must appeal to the hypotheses that were selected from the hypothesis space, not just the dichotomies they produced.

Let $H(\eta) = \{h \in \mathcal{H} \mid 2(\eta(i) - 1) = h(x_i) \text{ for all } x_i \in S\}$. A key observation is that $H(\eta)$ induces an equivalence relation on \mathcal{H} : $\mathcal{H} = \bigcup_{\eta \in M} H(\eta)$ and $H(\eta_1) \cap H(\eta_2) = \emptyset$ for any pair $\eta_1, \eta_2 \in M$. All hypotheses in each equivalence class, from the perspective of AdaBoost, are indistinguishable in the sense that picking any of them will result in no change in the trajectory of w_t . However, the weak learner might have a bias towards certain hypotheses in these classes. For example, perhaps the weak learner will always pick the “simplest” hypothesis in $H(\eta)$, based on some simplicity measure (depth of the tree, number of leaves, etc).

Regardless, in many cases the weak learner will always pick a specific hypothesis in each equivalence class. In this case, each class has a representative which we will denote h^η . Then whenever $\eta = \eta^{w_t}$ we have $h_t = h^\eta$ selected by the weak learner. We also get a finite number of hypothesis selection candidates in this case. If there are n rows in M , we are effectively reducing the hypothesis space from \mathcal{H} to $\{h^{\eta_1}, h^{\eta_2}, \dots, h^{\eta_n}\}$.

This is a common selection scheme in the case of decision stumps. In that case, a matrix very similar to M is often constructed. Then a pruning procedure is employed on M , removing repeated and dominated.² The result is a scheme following the above framework.

If we assume our weak learner follows the framework just described, we can extend our results in Theorem 16 to the whole space X .

2. A dichotomy η is dominated by η' if the set of examples incorrectly classified by η is a strict superset of the examples incorrectly classified by η' , because the error $\eta \cdot w_t > \eta' \cdot w_t$ for all t , thus η will never be selected by Optimal AdaBoost.

Theorem 18 *Suppose our weak learner follows the above framework. Then $\frac{1}{T}F_T(x)$ converges for all $x \in X$.*

Proof Let $\eta_t \equiv \eta^{w_t}$, then the hypothesis selected at iteration t can be represented as $h_t = h^{\eta_t}$. This yields

$$\begin{aligned} \frac{1}{T}F_T(x) &= \frac{1}{T} \sum_{t=1}^T \alpha_t h_t(x) \\ &= \frac{1}{T} \sum_{t=1}^T \alpha_t h^{\eta_t}(x) \\ &= \sum_{\eta \in M} \frac{1}{T} C_\eta(T) h^\eta(x). \end{aligned}$$

Where C_η is defined in the same way as in the proof of Theorem 16. As in the same proof, we have $\lim_{T \rightarrow \infty} \frac{1}{T} C_\eta(T) = C_\eta^*$. Hence,

$$\lim_{T \rightarrow \infty} \frac{1}{T} F_T(x) = \sum_{\eta \in M} C_\eta^* h^\eta(x).$$

■

Similarly, we can extend the convergence of the margin distribution to the whole space X .

Corollary 19 *Suppose our weak learner follows the above framework. Then $\text{margin}_T(x)$ converges for all $x \in X$.*

Proof We arrive at the convergence of $\lim_{T \rightarrow \infty} C_\eta(T) \left(\sum_{t=1}^T \alpha_t \right)^{-1} = \tilde{C}_\eta(T)$ the same way as in Corollary 17. Then, closely following the above proof, we get

$$\begin{aligned} \lim_{T \rightarrow \infty} \text{margin}_T(x) &= \lim_{T \rightarrow \infty} F_T(x) \left(\sum_{t=1}^T \alpha_t h_t(x) \right)^{-1} \\ &= \sum_{\eta \in M} \tilde{C}_\eta h^\eta(x). \end{aligned}$$

■

Recall that the full AdaBoost classifier is

$$H_T(x) = \text{sign}(F_T(x)).$$

In that equation, we can easily replace $\text{sign}(F_T(x))$ with $\text{sign}(\frac{1}{T}F_T(x))$. From the convergence of $\frac{1}{T}F_T(x)$, we would like to say that $H_T(x)$ converges as well. However, we have

a discontinuity in the sign function at 0. It may be the case that $\lim_{T \rightarrow \infty} \frac{1}{T} F_T(x) = 0$ for some $x \in X$, possibly yielding a non-existent limit for $\text{sign}(\frac{1}{T} F_T(x))$. In that case, $\lim_{T \rightarrow \infty} H_T(x)$ simply does not exist.

To overcome this obstacle, we will split the analysis. We will first consider the case that $\lim_{T \rightarrow \infty} \frac{1}{T} F_T(x) \neq 0$ almost everywhere;³ then we consider the case where this is not true.

For the first case, the limit of the classifier behaves nicely. This has the fascinating implication that the AdaBoost classifier itself, H_T , is converging in classification for almost all elements in the instance space X .

Theorem 20 *Suppose our weak learner follows the above framework, and that $\lim_{T \rightarrow \infty} \frac{1}{T} F_T(x) \neq 0$ for almost all $x \in X$. Then the limit $H^*(x) \equiv \lim_{T \rightarrow \infty} H_T(x)$ exists almost everywhere.*

Proof Let $H^*(x) \equiv \lim_{T \rightarrow \infty} H_T(x)$. Then the following holds:

$$\begin{aligned} H^*(x) &\equiv \lim_{T \rightarrow \infty} H_T(x) \\ &= \lim_{T \rightarrow \infty} \text{sign}\left(\frac{1}{T} F_T(x)\right) \\ &= \text{sign}(F^*(x)) \text{ (almost everywhere) .} \end{aligned}$$

■

If the AdaBoost classifier is converging in the limit, certainly its generalization error should as well. The generalization error of any 0-1 classifier $H(x)$ is written

$$\text{err}_D(H) = E_{(x,y) \in D} \left[\frac{1 - yH(x)}{2} \right] \quad (3)$$

It follows from the Lebesgue Dominated Convergence Theorem that the generalization error converges.

Theorem 21 *The limit of the generalization error, $\lim_{T \rightarrow \infty} \text{err}_D(H_T)$, exists.*

Proof First, we assume that each H_T is measurable on the probability space (D, Σ, P) . Otherwise it would not make sense to discuss generalization error in the first place, as it would not exist for our classifier at any time T . It follows that H^* is measurable, because $\lim_{T \rightarrow \infty} H_T(x) = H^*(x)$ almost everywhere. Furthermore, $\left(\frac{1 - yH_T(x)}{2}\right)$ is dominated by the constant function $f(x, y) = 1$ for all $(x, y) \in D$ and all T :

$$\left| \frac{1 - yH_T(x)}{2} \right| \leq 1.$$

Therefore, the conditions of Lebesgue Dominated Convergence Theorem are satisfied. We may then say that

3. If you let $F^*(x) = \lim_{T \rightarrow \infty} \frac{1}{T} F_T(x)$, intuitively we are saying that the decision boundary of the function F^* has measure 0 in the probability space (D, Σ, P) .

$$\begin{aligned}
 \lim_{T \rightarrow \infty} \text{err}_D(H_T) &= \lim_{T \rightarrow \infty} E_{(x,y) \in D} \left[\frac{1 - yH_T(x)}{2} \right] \\
 &= \lim_{T \rightarrow \infty} \int_{(x,y) \in D} \left(\frac{1 - yH_T(x)}{2} \right) dP \\
 &= \int_{(x,y) \in D} \lim_{T \rightarrow \infty} \left(\frac{1 - yH_T(x)}{2} \right) dP \\
 &= \int_{(x,y) \in D} \left(\frac{1 - yH^*(x)}{2} \right) dP
 \end{aligned}$$

■

We now consider the second case: the set of $x \in X$ such that $\lim_{T \rightarrow \infty} \frac{1}{T} F_T(x) = 0$ has non-zero measure. Let,

$$E = \left\{ (x, y) \in D \mid \lim_{T \rightarrow \infty} \frac{1}{T} F_T(x) \neq 0 \right\}.$$

We let the compliment of this set to take non-zero measure. Assuming that this set is measurable, we can say that the amount the generalization error can change in the limit depends on $P(D - E)$.

Theorem 22 *If E is measurable, the following are true:*

1. $\limsup_{T \rightarrow \infty} \text{err}_D(H_T) \leq \int_{(x,y) \in E} \left(\frac{1 - yH^*(x)}{2} \right) dP + P(D - E)$
2. $\liminf_{T \rightarrow \infty} \text{err}_D(H_T) \geq \int_{(x,y) \in E} \left(\frac{1 - yH^*(x)}{2} \right) dP - P(D - E)$

Additionally, if $\lim_{T \rightarrow \infty} \text{err}_D(H_T)$ exists, then

$$\left| \lim_{T \rightarrow \infty} \text{err}_D(H_T) - \int_{(x,y) \in E} \left(\frac{1 - yH^*(x)}{2} \right) dP \right| \leq P(D - E).$$

Proof

We can bound $\text{err}_D(H_T)$ as follows:

$$\begin{aligned}
 \text{err}_D(H_T) &= \int_{(x,y) \in D} \left(\frac{1 - yH_T(x)}{2} \right) dP \\
 &= \int_{(x,y) \in E} \left(\frac{1 - yH_T(x)}{2} \right) dP + \int_{(x,y) \in D-E} \left(\frac{1 - yH_T(x)}{2} \right) dP \\
 &\leq \int_{(x,y) \in E} \left(\frac{1 - yH_T(x)}{2} \right) dP + P(D - E).
 \end{aligned} \tag{4}$$

Symmetrically, we also have

$$\text{err}_D(H_T) \geq \int_{(x,y) \in E} \left(\frac{1 - yH_T(x)}{2} \right) dP - P(D - E). \quad (5)$$

We will consider only Equation 4, and results for Equation 5 will follow symmetrically. By taking \limsup on both sides of Equation 4, we see that

$$\begin{aligned} \limsup_{T \rightarrow \infty} \text{err}_D(H_T) &\leq \limsup_{T \rightarrow \infty} \int_{(x,y) \in E} \left(\frac{1 - yH_T(x)}{2} \right) dP + P(D - E) \\ &= \lim_{T \rightarrow \infty} \int_{(x,y) \in E} \left(\frac{1 - yH_T(x)}{2} \right) dP + P(D - E) \\ &= \int_{(x,y) \in E} \left(\frac{1 - yH^*(x)}{2} \right) dP + P(D - E). \end{aligned}$$

Symmetrically, we find

$$\liminf_{T \rightarrow \infty} \text{err}_D(H_T) \geq \int_{(x,y) \in E} \left(\frac{1 - yH^*(x)}{2} \right) dP - P(D - E).$$

Finally, if $\lim_{T \rightarrow \infty} \text{err}_D(H_T)$ exists, we have

$$\liminf_{T \rightarrow \infty} \text{err}_D(H_T) = \lim_{T \rightarrow \infty} \text{err}_D(H_T) = \limsup_{T \rightarrow \infty} \text{err}_D(H_T).$$

■

To recap, we conjectured that the dynamics of AdaBoost drift away from the discontinuities in its state space. Assuming this conjecture is true, we outlined a series of propositions that establish that AdaBoost's weight update \mathcal{A} , along with the functions representing the various secondary parameters, all satisfy the conditions for Birkhoff's Ergodic Theorem. Using Theorem 9 as our main tool, we derived convergence results for various aspects of AdaBoost, and showed the margin of all examples in the training set converge. Given that the weak learner follows a certain framework, the convergence results on the training set can be extended to the whole instance space X . Finally, if the decision boundary of F^* has probability 0, we conclude that the AdaBoost classifier H_T , along with the generalization error, converge. If the decision boundary has non-zero probability, we can say that the amount the generalization error can change in the limit depends on the probability of drawing an example on the decision boundary.

6. Characterizing the Inverse

When studying the dynamics of the AdaBoost update \mathcal{A} , it is natural to ask when given $w \in \Delta_m$, what is $\mathcal{A}^{-1}(w)$? Or similarly, when given $E \subset \Delta_m$ what is $\mathcal{A}^{-1}(E)$? To approach these questions, we decompose the inverse into a union of line segments.

First, let us simplify notation by decomposing a weight w into the components that η gets wrong and right.

Definition 23 Let (1) w_η^+ and (2) w_η^- be defined component-wise as

1. $w_\eta^+(i) = w(i) (1 - \eta(i))$
2. $w_\eta^-(i) = w(i) \eta(i)$.

The inverse set of \mathcal{T}_η at a point w is a line segment radiating outwards from w .

Proposition 24 Suppose we have $\eta \in M$ and $w \in \Omega$. Then $\mathcal{T}_\eta^{-1}(w) = \{2\rho w_\eta^- + 2(1 - \rho)w_\eta^+ \mid \rho \in [0, 1]\}$

Proof Let $L(\eta, w) = \{2\rho w_\eta^- + 2(1 - \rho)w_\eta^+ \mid \rho \in [0, 1]\}$. Consider an element $w' \in L(\eta, w)$. Clearly, $w' = 2\rho' w_\eta^- + 2(1 - \rho')w_\eta^+$ for some $\rho' \in [0, 1]$. Then,

$$\begin{aligned} \text{err}(\eta, w') &= \text{err}(\eta, 2\rho' w_\eta^-) + \text{err}(\eta, 2(1 - \rho')w_\eta^+) \\ &= 2\rho'(\eta \cdot w_\eta^-) \\ &= 2\rho'(\eta \cdot w) \\ &= (2\rho')\frac{1}{2} \\ &= \rho'. \end{aligned}$$

Using this fact, we see for i such that $\eta(i) = 1$

$$\begin{aligned} [\mathcal{T}_\eta(w')]_i &= w'_i \times \frac{1}{2(\eta \cdot w')} \\ &= 2\rho' w_i \times \frac{1}{2\rho'} \\ &= w_i. \end{aligned}$$

And similarly, for i such that $\eta(i) = 0$

$$\begin{aligned} [\mathcal{T}_\eta(w')]_i &= w'_i \times \frac{1}{2(1 - (\eta \cdot w'))} \\ &= 2(1 - \rho')w_i \times \frac{1}{2(1 - \rho')} \\ &= w_i. \end{aligned}$$

Pulling the cases together, we conclude that $\mathcal{T}_\eta(w') = w$ and $w' \in \mathcal{T}_\eta^{-1}(w)$.

Instead, suppose that $w' \in \mathcal{T}_\eta^{-1}(w)$. From Definition 5, we see that

$$w'_i = \begin{cases} 2(\eta \cdot w')w_i & \text{if } \eta(i) = 1 \\ 2(1 - (\eta \cdot w'))w_i & \text{if } \eta(i) = 0 \end{cases}.$$

Setting $\rho' = \eta \cdot w'$, we see that $w' \in L(\eta, w)$. ■

So, $\mathcal{T}_\eta(w)$ has a very clean inverse, being simply a line through simplex space. But, it is important to note that $\mathcal{T}_\eta(w)$ is hypothetical, asking “where would w go if $\eta = \eta^w$?” and is not the true AdaBoost weight update, $\mathcal{A}(w)$. Though, the inverse $\mathcal{A}^{-1}(w)$ does decompose into a union of these line segments.

Proposition 25 *Let $w \in \Omega$. Then $\mathcal{A}^{-1}(w) = \bigcup_{\eta \in M} (\mathcal{T}_\eta^{-1}(w) \cap \sigma^*(\eta))$.*

Proof Take $w' \in \mathcal{A}^{-1}(w)$. Then, $w' \in \sigma(\eta^w)$ and $w' \in \mathcal{T}_{\eta^w}^{-1}(w)$. Therefore, $w' \in \bigcup_{\eta \in M} (\mathcal{T}_\eta^{-1}(w) \cap \sigma^*(\eta))$.

Instead, take $w' \in \bigcup_{\eta \in M} (\mathcal{T}_\eta^{-1}(w) \cap \sigma^*(\eta))$. It must be the case that $w' \in \mathcal{T}_{\eta^w}^{-1}(w) \cap \sigma^*(\eta^w)$, because w' can only be in $\sigma^*(\eta')$ for one possible $\eta' \in M$, namely $\eta' = \eta^w$. But by Definition 1, we see that implies $\mathcal{A}(w') = w$. Therefore, $w' \in \mathcal{A}^{-1}(w)$. ■

7. Satisfying Birkhoff’s Ergodic Theorem

This section is devoted to proving Proposition 14 and 15, and to do so we will use the fundamental Krylov-Bogolyubov theorem, listed as 26 below. For a given dynamical system that meets certain conditions, Krylov-Bogolyubov tells us that the system is *measure-preserving* on some Borel probability measure. We will apply this theorem on our trapped attracting set Ω_∞ to show that \mathcal{A} admits an invariant measure.

Recall that part (2) of Assumption 12 says that if $\sigma(\eta) \cap \sigma(\eta') \cap G \neq \emptyset$, then $\eta = \eta'$ with respect to the weights in G . Let $\mathcal{H}_1, \mathcal{H}_2, \dots, \mathcal{H}_k$ be the equivalence classes over the rows of M where, for every pair $\eta, \eta' \in H_i$, we have $\sum_{i: \eta(i) \neq \eta'(i)} w(i) = 0$ for all $w \in G$. We construct a new matrix M' where row i is $\text{AdaSelect}(\mathcal{H}_i)$. On G , it is easy to see that the dynamics of \mathcal{A} with respect to M is the same with respect to M' . Meanwhile, \mathcal{A} with respect to M' is bounded away from ties. Whenever part (2) of Assumption 12 holds, we will restrict our analysis to \mathcal{A} with respect to M' , but for simplicity keep the notation M .

A couple of concepts are essential in understanding this theorem. First, Krylov-Bogolyubov requires that we deal with a system of the form (X, T) , where X is the state space and T a *topology* on it. Furthermore, (X, T) needs to be *metrizable*, meaning the topology T can be induced by some metric. To simplify matters, we will treat Δ_m as a metric space with the metric

$$d(w_1, w_2) = \sum_{i=1}^m |w_1(i) - w_2(i)|.$$

We will not use d directly in our proofs; but when we discuss convergence of sequences of weights in Δ_m , we will implicitly use the metric. ⁴

-
4. Convergence is meaningless without such a metric. The definition of closed and open sets also implicitly use d : closed sets are the sets in Δ_m that contain all of their limit points. That is, a set E is closed if, given any convergent sequence $\{w_i\} \subseteq E$, we have $\lim_{i \rightarrow \infty} w_i \in E$. Compact sets are closed sets that are bounded. We are only considering subsets of Δ_m , so all such subsets are bounded and any closed subset will be compact.

Krylov-Bogolyubov requires that the state space is compact. We want to apply the theorem on Ω_∞ , so this is the set we must scrutinize. As mentioned before, this set is contained within Δ_m , so we know it is bounded. Hence, we only need to show that Ω_∞ is closed. That is the motivation behind Theorem 28.

Additionally, Krylov-Bogolyubov requires that \mathcal{A} is continuous on Ω_∞ . We have stated in Assumption 12 that Ω_∞ is bounded away from type 1 discontinuities, so what remains to show is that it does not contain type 2 discontinuities. This is accomplished in Theorem 30.

Theorem 26 (*Krylov-Bogolyubov*) *Let (X, T) be a compact, metrizable topological space and $F : X \rightarrow X$ a continuous map. Then F admits an invariant Borel probability measure.*

We will begin by showing the compactness of Ω_∞ , given Assumption 12. We first approach this by proving the following lemma, which states that any limit point $w \in \Omega_\infty$ has a corresponding limit point $w' \in G$ such that $\mathcal{A}(w') = w$.

Lemma 27 *Suppose Assumption 12 holds. Let $\{w_i\}$ be an arbitrary convergent sequence in Ω_∞ , and call its limit w . Then there exists a second convergent sequence $\{w'_i\} \subset \Omega$, such that $\mathcal{A}(\lim_{i \rightarrow \infty} w'_i) = w$.*

Proof Let $\{w_i\} \subset \Omega_\infty$ be such a sequence as described in the hypothesis, and let $w = \lim_{i \rightarrow \infty} w_i$. From the compactness of G , we have $w \in G$. Additionally, as $w_i \in \Omega_\infty$, there must exist a $w'_i \in \Omega_\infty$ such that

$$w_i = \mathcal{A}(w'_i). \quad (6)$$

Let $\{w'_i\} \subset \Omega_\infty$ be a sequence composed of such elements. We will now proceed to show that there exists a subsequence of $\{w'_i\}$ that has a limit $w' \in \mathcal{A}^{-1}(w)$.

Consider subsets of G the form

$$P^*(\eta) = \{g \in G \mid g \in \sigma^*(\eta)\}.$$

Note that G does not contain any elements that are tied, hence we have

$$G = \bigcup_{\eta \in M} P^*(\eta).$$

There exists an $\eta \in M$ such that $P^*(\eta)$ contains infinite elements from the sequence $\{w'_i\}$. Let $\{w'_{i_j}\}$ be the subsequence of $\{w'_i\}$ that is contained in $P^*(\eta)$. Note that G is sequentially compact because G is compact subset of a metric space. Therefore, there exists a convergent subsequence $\{w'_{i_{j_k}}\}$, and call its limit w' . We claim that $w' \in P^*(\eta)$.

Let $P(\eta) = \{g \in G \mid g \in \sigma(\eta)\}$, which is simply the intersection $G \cap \sigma(h)$ and closure of $P^*(\eta)$. It follows that $P(\eta)$ is closed, because both sets involved in the intersection are closed. Also note that $P^*(\eta) \subseteq P(\eta)$. The sequence $\{w'_{i_{j_k}}\}$ is therefore contained in $P(\eta)$, yielding $w' \in P(\eta)$. Now, either $w' \in P^*(\eta)$ or $w' \in P(\eta) - P^*(\eta)$, the later containing only weights in which η is tied with another row in M . The second case is impossible because the hypothesis of the theorem does not allow ties, so we must conclude $w' \in P^*(\eta)$.

Now we proceed to show that $w' = \mathcal{A}(w)$. From Equation 6, it is clear that there is a subsequence $\{w_{i_{j_k}}\}$ of $\{w_i\}$ such that $w_{i_{j_k}} = \mathcal{A}(w'_{i_{j_k}})$. Whereby,

$$\begin{aligned}
 \lim_{k \rightarrow \infty} \mathcal{A}(w'_{i_{j_k}}) &= \lim_{k \rightarrow \infty} w_{i_{j_k}} \\
 &= \lim_{i \rightarrow \infty} w_i \\
 &= w.
 \end{aligned} \tag{7}$$

By continuity of \mathcal{A} on G (see Theorem 11), it is clear that

$$\begin{aligned}
 \lim_{k \rightarrow \infty} \mathcal{A}(w'_{i_{j_k}}) &= \mathcal{A}\left(\lim_{k \rightarrow \infty} w'_{i_{j_k}}\right) \\
 &= \mathcal{A}(w').
 \end{aligned} \tag{8}$$

Then, combining Equation 7 and Equation 8, we conclude that $\mathcal{A}(w') = w$. \blacksquare

Given any limit point w of Ω_∞ , this lemma lets us construct an infinite orbit backwards from w contained entirely in G , whereby $w \in \Omega_\infty$, giving us compactness. This is formalized in the next theorem.

Theorem 28 *Suppose Assumption 12 holds. Then, Ω_∞ is compact.*

Proof Let $\{w_i\}$ be an arbitrary convergent sequence contained in Ω_∞ , and let $w = \lim_{i \rightarrow \infty} w_i$. By Lemma 27, there exists a sequence $\{w_{k(1)}^{(1)}\} \subset \Omega_\infty$ converging to $w^{(1)} \in G$ such that $\mathcal{A}(w^{(1)}) = w$. However, notice that $\{w_{k(1)}^{(1)}\}$ also satisfies the hypothesis of Lemma 27. Applying the lemma to $\{w_{k(1)}^{(1)}\}$, we get $\{w_{k(2)}^{(2)}\} \subset \Omega_\infty$ converging to $w^{(2)} \in G$ such that $\mathcal{A}(w^{(2)}) = w^{(1)}$, therefore $\mathcal{A}^{(2)}(w^{(2)}) = w$. We can continue in this way to generate $w^{(n)} \in G$ such that $\mathcal{A}^{(n)}(w^{(n)}) = w$ for any n . Therefore, $w \in \mathcal{A}^{(n)}(\Delta_m)$ for all $n \in \mathbf{N}$ and we must conclude that $w \in \Omega_\infty$. Because w was the limit of an arbitrary convergent sequence of Ω_∞ , it must be the case that Ω_∞ is compact. \blacksquare

Now we must turn to understanding the continuity properties of \mathcal{A} . As previously mentioned, \mathcal{A} is continuous on most points in its state space. Assumption 12 tells us that Ω_∞ is bounded away from type 1 discontinuities. But, if \mathcal{A} is continuous on our attracting set Ω_∞ , we must show that Ω_∞ contains no points that have a row with error zero.

The following lemma takes a step towards this goal. It shows that, given a point $w \in \Omega(\eta)$, if the error of a hypothesis in M is low on w , then the error of that same hypothesis on the inverse of w is not too large. Not only that, but the error of η on the inverse also is not too large. This lemma is used in proving the next theorem, which tells us that AdaBoost is bounded away from type 2 discontinuities.

Lemma 29 *Let $\eta, \eta' \in M$ and $w \in \Omega(\eta)$. If $\eta' \cdot w \leq \epsilon_0$, then for all $w' \in \mathcal{A}^{-1}(w)$ we have $\eta' \cdot w' \leq 2\epsilon_0$ and $\eta \cdot w' \leq 2\epsilon_0$*

Proof Let $M_{\frac{1}{2}}(w) = \{\eta \in M \mid \eta \cdot w = \frac{1}{2}\}$, and let $w(\rho, \eta) = 2\rho w_{\eta}^{-} + 2(1 - \rho)w_{\eta}^{+}$. Let $L = \{w(\rho, \eta) \mid \eta \in M_{\frac{1}{2}}(w), \rho \in [0, \frac{1}{2}]\}$, and note that $\mathcal{A}^{-1}(w) \subseteq L$, so it suffices to show that the lemma holds for all elements in L .

Take an arbitrary pair $\eta \in M_{\frac{1}{2}}(w)$, $\eta' \in M$, and $\rho \in [0, \frac{1}{2})$. We can decompose $\eta' \cdot w(\rho, \eta)$ as

$$\eta' \cdot w(\rho, \eta) = 2\rho(\eta' \cdot w_{\eta}^{-} - \eta' \cdot w_{\eta}^{+}) + 2(\eta' \cdot w_{\eta}^{+}). \quad (9)$$

To upper bound $\eta' \cdot w(\rho, \eta)$, we consider three cases depending on the relationship between $\eta' \cdot w_{\eta}^{-}$ and $\eta' \cdot w_{\eta}^{+}$.

1. If $\eta' \cdot w_{\eta}^{-} > \eta' \cdot w_{\eta}^{+}$, then

$$\eta' \cdot w(\rho, \eta) < \eta' \cdot w_{\eta}^{-} + \eta' \cdot w_{\eta}^{+} = \eta' \cdot w \leq \epsilon_0.$$

2. If $\eta' \cdot w_{\eta}^{-} \leq \eta' \cdot w_{\eta}^{+}$, then

$$\eta' \cdot w(\rho, \eta) \leq 2\eta' \cdot w_{\eta}^{+} \leq 2(\eta' \cdot w) \leq 2\epsilon_0.$$

Taking the largest of the upper bounds, we conclude that $\eta' \cdot w(\rho, \eta) \leq 2\epsilon_0$. Now, if $w(\rho, \eta) \in \mathcal{A}^{-1}(w)$, it follows that $h = \text{AdaSelect}(w(\rho, \eta))$. Therefore $h \cdot w(\rho, \eta) \leq \eta' \cdot w(\rho, \eta) \leq 2\epsilon_0$. \blacksquare

We can now apply the above lemma in a recursive manner to show that there exists an n_0 such that for all $n > n_0$ the error of any hypothesis on the points in $\mathcal{A}^{(n)}(\Delta_m)$ is bounded away from zero.

Theorem 30 *There exists an $\epsilon_* > 0$ and n_0 such that, for all $n > n_0$ we have $\eta \cdot w \geq \epsilon_*$ for all $w \in \mathcal{A}^{(n)}(\Delta_m)$ and $\eta \in M$.*

Proof Set $n_0 = |M| + 1$, and $\epsilon_* < \frac{1}{2^{n_0}}$. Take an arbitrary $\eta \in M$ and $w \in \Delta_m$ such that $\eta \cdot w < \epsilon_*$. We will show that such a w is not contained within $\mathcal{A}^{(n)}$ for $n > n_0$.

Let $w^{(1)} \in \mathcal{A}^{-1}(w)$. If no such $w^{(1)}$ exists, we have already demonstrated our goal. Then, let $\eta^{(1)} = \eta^{w^{(1)}}$. By Lemma 29, we know that $\eta^{(1)} \cdot w^{(1)} \leq 2\epsilon_*$. Continuing in this way, let $w^{(2)} \in \mathcal{A}^{-1}(w^{(1)})$, which we can assume exists by the same argument made for $w^{(1)}$. Let $\eta^{(2)} = \eta^{w^{(2)}}$, and note that $\eta^{(2)} \neq \eta^{(i)}$ for $i < 2$ because

$$\begin{aligned} \eta^{(2)} \cdot w^{(1)} &= \frac{1}{2} \\ &> 2\epsilon_* \\ &\geq \eta^{(1)} \cdot w^{(1)}. \end{aligned}$$

We can continue this template out to n_0 . Let $w^{(n_0)} \in \mathcal{A}^{-1}(w^{(n_0-1)})$. We claim that such a $w^{(n_0)}$ cannot exist. For sake of contradiction, suppose it did. Then, let $\eta^{(n_0)} = \eta^{w^{(n_0)}}$.

From Lemma 29, we know that $\eta^{(n_0)} \neq \eta^{(i)}$ for all $i < n_0$ because

$$\begin{aligned} \eta^{(n_0)} \cdot w^{(n_0-1)} &= \frac{1}{2} \\ &> 2^{(n_0-1)} \epsilon_* \\ &\geq \eta^{(i)} \cdot w^{(n_0-1)}. \end{aligned}$$

Furthermore, all $\eta^{(i)}$ are unique by the construction of this sequence. Because $n_0 = |M| + 1$, the sequence $\{\eta^{(1)}, \eta^{(2)}, \dots, \eta^{(n_0-1)}\} = M$. But because $\eta^{(n_0)} \neq \eta^{(i)}$ for all $i < n_0$, $\eta^{(n_0)}$ is not in M . As this is a contradiction, we must conclude that no such $w^{(n_0)}$ exists, and that $\mathcal{A}^{-1}(w^{(n_0-1)}) = \emptyset$. Our selection of $w^{(i)}$ was arbitrary in each step, so we can also conclude that there does not exist any w' such that $A^{(n_0)}(w') = w$, or else it would have been reached by the above procedure. Finally, this shows that w is not contained in $\mathcal{A}^{(n_0)}(\Delta_m)$. ■

We now have the tools necessary to prove Proposition 14, originally stated in Section 5. **Proof** (Proof of Proposition 14) If the condition holds, then Ω_∞ is a compact and metrizable topological space. Also, \mathcal{A} is continuous on Ω_∞ . It follows from Krylov-Bogolyubov that there exists a Borel probability measure μ that is invariant on \mathcal{A} . We may extend this measure to all of Δ_m . Let $A \subseteq \Delta_m$, then define $\mu(A) = \mu(A \cap \Omega_\infty)$. ■

Also, we now have the tools necessary to prove Proposition 15, stated in the same section. We prove a slightly modified one tailored to our new context.

Proof (Proof of Proposition 15)

1. Because $\epsilon(w)$ is the minimum of a finite set of continuous functions, it follows that $\epsilon(w)$ is continuous as well. In the case of a Borel algebra, continuity implies measurability. It follows that

$$\int_{w \in \Omega_\infty} |\epsilon(w)| d\mu < \int_{w \in \Omega_\infty} \frac{1}{2} d\mu < \frac{1}{2}.$$

Therefore, $\epsilon(w) \in L^1(\mu)$.

2. Because $\epsilon(w)$ is continuous and bounded away from 0 on Ω_∞ , it follows that $\alpha(w)$ is continuous as well. As above, this implies measurability. From $\epsilon(w) > \epsilon_* > 0$, we have an upper bound on $\alpha(w)$ we will call α^* . Therefore, we then have

$$\int_{w \in \Omega_\infty} |\alpha(w)| d\mu \leq \int_{w \in \Omega_\infty} \alpha^* d\mu = \alpha^* \mu(\Omega_\infty) = \alpha^*.$$

Whereby $\alpha(w) \in L^1(\mu)$.

3. $\chi_{\sigma^*(\eta)}(w)$ is measurable and bounded above by 1. Therefore, it is in $L^1(\mu)$. ■

8. Preliminary Experimental Results Support the No-Ties Conjecture

This section discusses *preliminary* experimental evidence that Assumption 12 holds in practice. Recall that the assumption requires that, for any pair $\eta, \eta' \in M$, either (1) there are no ties between η and η' in the limit, or (2) if they are tied, they are effectively the same with respect to the weights in the limit. We provide empirical evidence on a few commonly used data sets in practice suggesting that these two conditions seem to be satisfied.

In Figure 3, we AdaBoost decision stumps on the Heart-Disease (Frank and Asuncion, 2010b), Sonar (Frank and Asuncion, 2010a), and Breast-Cancer (Frank and Asuncion, 2010a) datasets, while tracking the difference between the error of the best row and the second best row of M at each round t . Let η be the optimal row in M at round t . When looking for the second best row at round t , we ignore rows η' such that $\sum_{i:\eta(i) \neq \eta'(i)} w_t(i) < \epsilon$, where we set $\epsilon = 10^{-15}$ for Heart-Disease and Sonar, and $\epsilon = 10^{-10}$ for Breast-Cancer. We start AdaBoost from an initial weight drawn uniformly at random from the m -simplex (as opposed to the traditionally used weight corresponding to a uniform distribution over the training examples).

The difference between best and second best row tends to decrease to ϵ early on. This happens because some weights go to zero for non-minimal-margin examples, which from now on we refer to as the “support vectors” because of their similar interpretation to those examples in SVMs.⁵ Such zero-weight examples could cause certain rows to become essentially equal with respect to the weights. Once such weights go below ϵ , a condition which we equate to essentially satisfying part (2) of Assumption 12, we ignore these “equivalent” rows. In turn, this causes the trajectory of the differences between best and second best to jump upwards. After a sufficient number of rounds, the set of support vectors manifests, and this jumping behavior stops. At this point, it appears the distance from ties is bounded away from zero, suggesting Assumption 12 holds for AdaBoosting decision stumps on these datasets.

Figure 4 provides reasonably clear evidence for the convergence of the Optimal AdaBoost classifier when boosting decision stumps on the Breast-Cancer dataset ???. In this figure, the margin for every example seems to be converging: from rounds 90k to 100k there is very little change, as seen most clearly in plot (b) of that figure. Figure 5 shows convergence of the minimum margin; this is essentially a more complete view of the convergence of the minimum margin clearly seen in the histograms in Figure 4(c). This converging behavior is as predicted by the theoretical work in Section 5.

5. For all training examples i , denote by $\beta_T(i) \equiv y_i \text{margin}_T(x_i)$ the “signed” margin of example i . From our convergence results we can show that $\beta^{\min} \equiv \lim_{T \rightarrow \infty} \min_i \beta_T(i)$ exists. We can also show that $\beta^{\min} = \lim_{T \rightarrow \infty} \sum_i w_{T+1}(i) \beta_T(i)$. This implies that, for all training examples i , $\lim_{T \rightarrow \infty} \beta_T(i) > \beta^{\min}$ implies $\lim_{T \rightarrow \infty} w_{T+1}(i) = 0$; and that $\lim_{T \rightarrow \infty} w_{T+1}(i) > 0$ implies $\lim_{T \rightarrow \infty} \beta_T(i) = \beta^{\min}$. Also, assuming training examples with different outputs, there always exists a pair of different-label examples (j, k) , with $y_j = 1$ (positive example) and $y_k = -1$ (negative example), such that $\lim_{T \rightarrow \infty} w_{T+1}(j) > 0$ and $\lim_{T \rightarrow \infty} w_{T+1}(k) > 0$ (because the error $\eta_T \cdot w_{T+1} = \frac{1}{2}$, where η_T is the row of M corresponding to the dichotomy selected at round T). This in turn implies $\lim_{T \rightarrow \infty} \beta_T(j) = \lim_{T \rightarrow \infty} \beta_T(k) = \beta^{\min}$, leading to our interpretation of the set $\{i \mid \lim_{T \rightarrow \infty} \beta_T(i) = \beta^{\min}\}$ as the set of “support vectors.”

9. Closing Remarks

In this paper we provide a non-constructive proof of existence of an invariant measure, provided the dynamics of AdaBoost satisfy certain conditions. An improvement of this result would be a direct proof for the existence of such a measure, perhaps by proving that AdaBoost always satisfies the conditions. An even stronger result would be a constructive proof of such a measure. There are some hints for such a proof lying in the simple nature of the inverse of the AdaBoost mapping \mathcal{A} .

While we provide convergence proofs, we do not provide convergence *rates*. We suspect that the rate varies significantly between datasets and choice of weak learner. For example, on datasets where AdaBoost tends to overfit, we suspect the rate of convergence is slower. On the other hand, the stronger the weak learner, the quicker the rate of convergence seems to be. Hence why the generalization error of AdaBoost seems to quickly converge when using decision trees.

Finally, we would like to say something about the *quality* of the generalization error, beyond just that it converges. In all of our experiments involving decision stumps, we have observed a logarithmic growth of the number of *unique* hypothesis contained in the combined AdaBoost classifier as a function of time. Such a logarithmic growth yields a tighter data-dependent bound on the generalization of the AdaBoost classifier. We believe that the distribution of the invariant measure over the regions $\sigma(h)$ is an important factor for this behavior: the relative frequency of selecting each hypothesis seems Gamma distributed.

References

- Peter L. Bartlett and Mikhail Traskin. Adaboost is consistent. *J. Mach. Learn. Res.*, 8:2347–2368, December 2007. URL <http://dl.acm.org/citation.cfm?id=1314498.1314574>.
- Leo Breiman. Arcing classifiers. *The Annals of Statistics*, 26(3):801–849, 1998.
- Leo Breiman. Prediction games and arcing algorithms. *Neural Comput.*, 11(7):1493–1517, October 1999. URL <http://dx.doi.org/10.1162/089976699300016106>.
- Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. <http://dx.doi.org/10.1023/A:1010933404324>.
- Leo Breiman. Machine learning. Wald Lecture speech presented at the 277th meeting of the Institute of Mathematical Statistics, held in Banff, Alberta, Canada, July 2002. <http://www.stat.berkeley.edu/~breiman/wald2002-1.pdf>.
- Michael Collins, Robert E. Schapire, and Yoram Singer. Logistic regression, AdaBoost and Bregman distances. In *Machine Learning*, volume 48, pages 253–285. Kluwer Academic Publishers, 2002.
- Harris Drucker and Corinna Cortes. Boosting decision trees. In *NIPS*, pages 479–485, 1995.
- A. Frank and A. Asuncion. UCI machine learning repository, 2010a. URL <http://archive.ics.uci.edu/ml>.

- A. Frank and A. Asuncion. UCI machine learning repository, 2010b. URL <http://archive.ics.uci.edu/ml>. Heart Disease Dataset, PI Data Collectors: Andras Janosi, M.D. (Hungarian Institute of Cardiology, Budapest), William Steinbrunn, M.D. (University Hospital, Zurich, Switzerland), Matthias Pfisterer, M.D. (University Hospital, Basel, Switzerland) and Robert Detrano, M.D., Ph.D. (V.A. Medical Center, Long Beach and Cleveland Clinic Foundation).
- Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 325–332, 1996.
- Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119 – 139, 1997. URL <http://www.sciencedirect.com/science/article/pii/S002200009791504X>.
- Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28:2000, 1998.
- Adam J. Grove and Dale Schuurmans. Boosting in the limit: Maximizing the margin of learned ensembles. In *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*, AAAI '98/IAAI '98, pages 692–699, Menlo Park, CA, USA, 1998. American Association for Artificial Intelligence. URL <http://dl.acm.org/citation.cfm?id=295240.295766>.
- Aur lie C. Lozano, Sanjeev R. Kulkarni, and Robert E. Schapire. Convergence and consistency of regularized boosting algorithms with stationary beta-mixing observations. In *Advances in Neural Information Processing Systems*, volume 18, 2006.
- G bor Lugosi and Nicolas Vayatis. On the bayes-risk consistency of regularized boosting methods. *The Annals of Statistics*, 32(1):30–55, 2004.
- Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. Boosting algorithms as gradient descent. In *In Advances in Neural Information Processing Systems 12*, pages 512–518. MIT Press, 2000.
- David Mease and Abraham Wyner. Evidence contrary to the statistical view of boosting. *J. Mach. Learn. Res.*, 9:131–156, June 2008. URL <http://dl.acm.org/citation.cfm?id=1390681.1390687>.
- Stefano Merler, Bruno Caprile, and Cesare Furlanello. Parallelizing AdaBoost by weights dynamics. *Computational Statistics and Data Analysis*, 51(5):2487 – 2498, 2007. URL <http://www.sciencedirect.com/science/article/pii/S0167947306003239>.
- Indraneel Mukherjee, Cynthia Rudin, and Robert E. Schapire. The rate of convergence of Adaboost. *Journal of Machine Learning Research - Proceedings Track*, 19:537–558, 2011.
- Nikunj Chandrakant Oza. *Online Ensemble Learning*. Ph.d. dissertation, Graduate Division, University of California, Berkeley, Department of Electrical Engineering and Computer Science, Berkeley, CA, Fall 2001.

- Alon Zakai Peter J. Bickel, Ya'acov Ritov. Some theory for generalized boosting algorithms. *Journal of Machine Learning Research*, 7:705–732, May 2006.
- J. R. Quinlan. Bagging, boosting, and c4.5. In *In Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 725–730. AAAI Press, 1996.
- Saharon Rosset, Ji Zhu, and Trevor Hastie. Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5:941–973, August 2004.
- Cynthia Rudin, Ingrid Daubechies, and Robert E. Schapire. The dynamics of AdaBoost: Cyclic behavior and convergence of margins. *Journal of Machine Learning Research*, 5: 1557–1595, 2004.
- Cynthia Rudin, Robert E. Schapire, and Ingrid Daubechies. Precise statements of convergence for AdaBoost and arc-gv. In *AMS-IMS-SIAM Joint Summer Research Conference on Machine and Statistical Learning, Prediction and Discovery*, pages 131–145, 2007.
- Cynthia Rudin, Robert E. Schapire, and Ingrid Daubechies. Open problem: Does AdaBoost always cycle? *Journal of Machine Learning Research - Proceedings Track*, 23:46.1–46.4, 2012.
- Robert E. Schapire and Yoav Freund. *Boosting: Foundations and Algorithms*. MIT Press, 2012.
- Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26 (3):824–832, 1998.
- Matus Telgarsky. A primal-dual convergence analysis of boosting. *J. Mach. Learn. Res.*, 13:561–606, March 2012. <http://dl.acm.org/citation.cfm?id=2188385.2188405>.
- Yongxin Taylor Xi, Zhen James Xiang, Peter J. Ramadge, and Robert E. Schapire. Speed and sparsity of regularized boosting. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, 2009.
- Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.
- Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33(4):1538–1579, 2005.

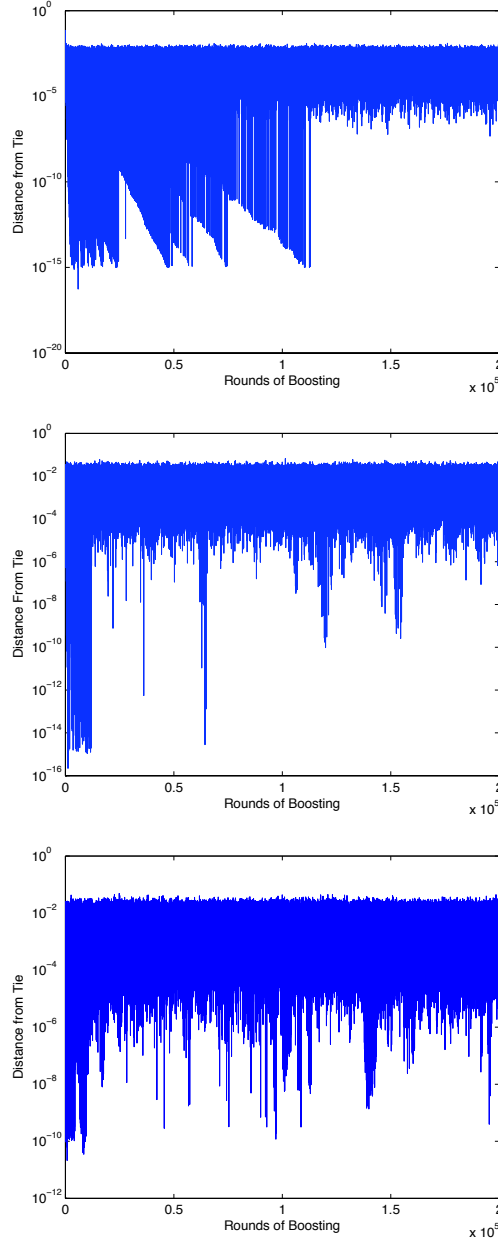


Figure 3: **Evidence for No Ties.** These plots depict the difference between the errors for the best row and the second best row (log scale) as a function of the number of rounds of boosting decision stumps on the Heart-Disease (Frank and Asuncion, 2010b) (**top**), Sonar (Frank and Asuncion, 2010a) (**center**), and Breast-Cancer Frank and Asuncion (2010a) datasets (**bottom**). The behavior depicted in these plots suggests that Assumption 12 holds. Recall that, as described in the body of the text, when looking for the second best row at time t , we ignore rows η' such that $\sum_{i:\eta(i) \neq \eta'(i)} w_t(i) < \epsilon$, where we set $\epsilon = 10^{-15}$ for Heart-Disease and Sonar, and $\epsilon = 10^{-10}$ for Breast-Cancer, and that we start AdaBoost from a weight over the training examples drawn uniformly at random from the m -simplex.

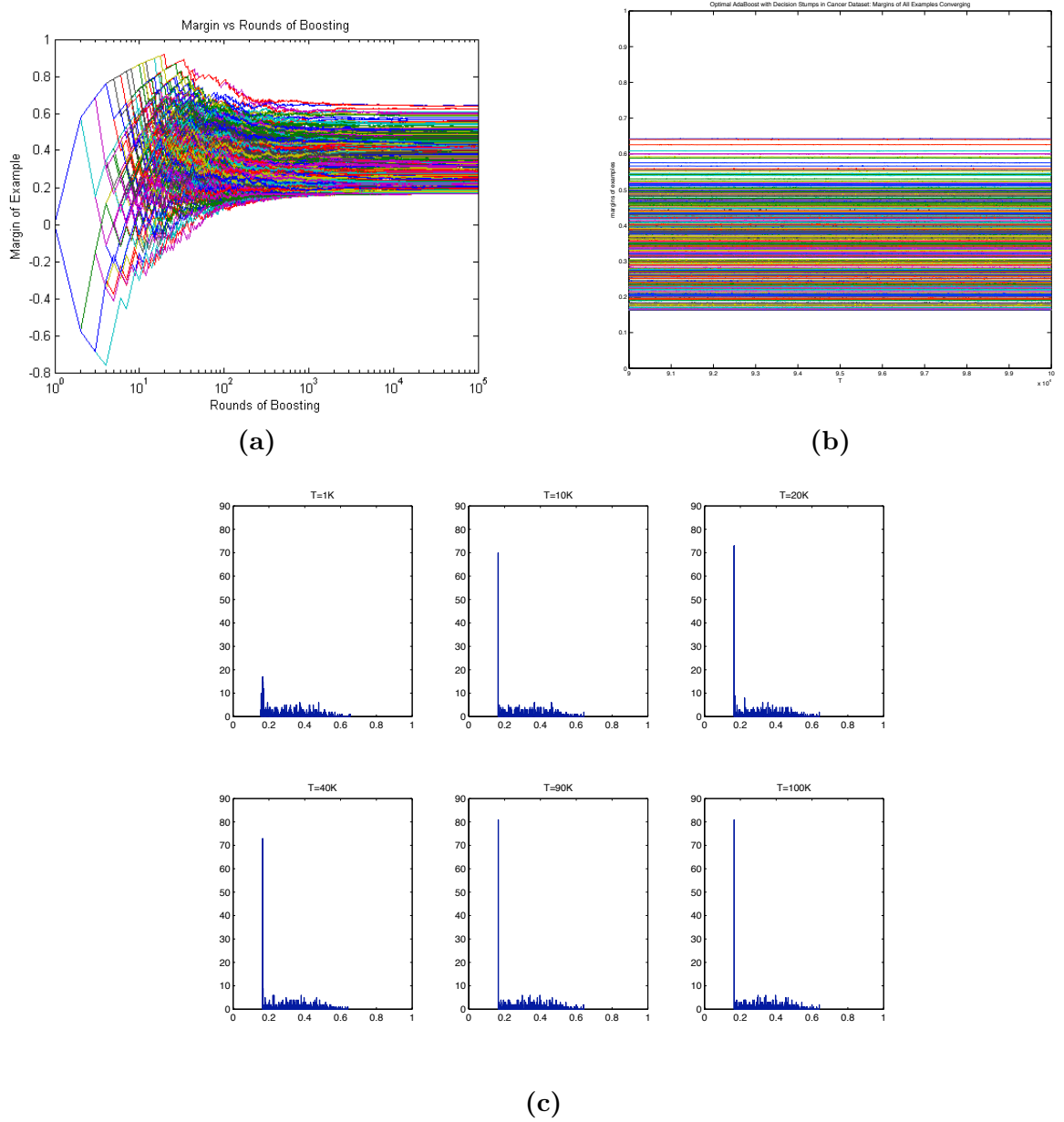


Figure 4: **Evidence for the Convergence of Optimal AdaBoost Classifier when Boosting Decision Stumps on the Breast-Cancer Dataset.** Plot (a) shows the behavior of the “signed” margin $y_i \text{ margin}_T(x_i)$ of every example i as a function of the number of rounds T of boosting (log scale). Plot (b) is a closer look at the asymptotic behavior of these margins from rounds $T = 90K$ to $100K$. Evidence for the convergence of the signed margins is more evident at this resolution. Plot (c) shows the histogram of signed margins at rounds $T = 1K, 10K, 20K, 40K, 90K, 100K$. The histograms contain 200 bins. Note that they are all positive because from the theory of AdaBoost, assuming the weak-learning hypothesis holds, all the training examples are correctly classified eventually after some finite number of rounds (logarithmic in m), so that the signed margin will always be positive. Note also that the examples in the histogram whose signed margin is closest to zero correspond to the “support vectors” (see main text for further discussion).

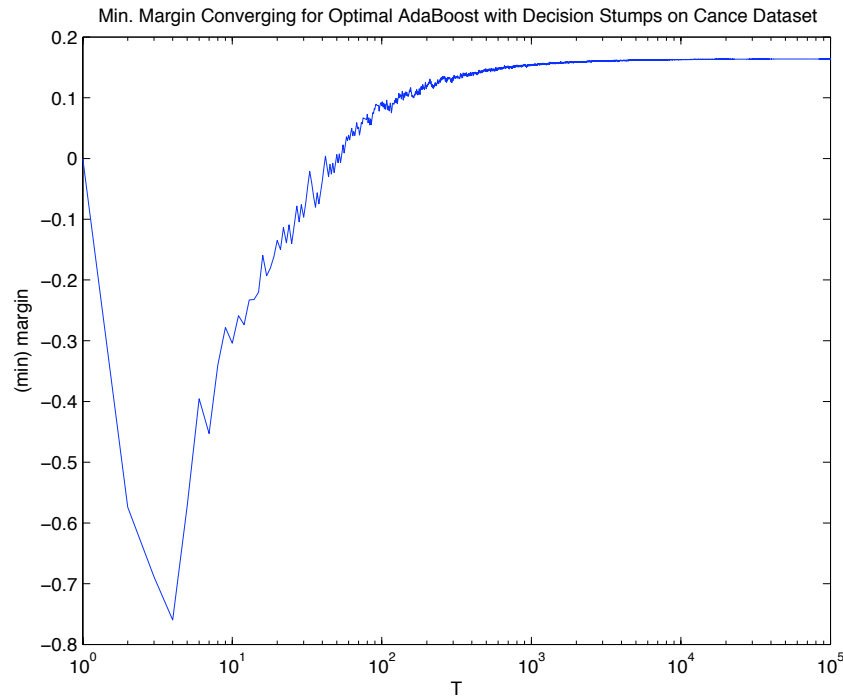


Figure 5: **Evidence for the Convergence of the Minimum Margin.** This plot depicts the minimum margin as a function of the number of rounds of boosting (log scale) on the Breast-Cancer dataset (Frank and Asuncion, 2010a), using decision stumps. This is an isolation of the minimum margin from Figure 4(c).